自然語言處理 HW1

成功大學114資訊系 F74102022唐文蔚

|  |  |  |
| --- | --- | --- |
| Environment | Colab | Ubuntu 20.04.6 LTS |
| CPU | Colab | Intel(R) Xeon(R) Gold 6242R CPU @ 3.10GHz |
| Python Version | Colab | Python 3.9.20 |

* Pre-processing (TODO 5)

(v1)照著投影片的方式逐一尋找資料，實作Pre-process步驟。

使用的方法包括：

1. 移除非英文字
2. 移除stop words：<https://www.geeksforgeeks.org/removing-stop-words-nltk-python/>
3. Lemmatization：<https://www.geeksforgeeks.org/python-lemmatization-with-nltk/>
4. 篩選頻率較高字

一張含有 文字, 螢幕擷取畫面, 陳列, 電腦 的圖片

自動產生的描述

(v2) 由於v1正確率極低，因此嘗試用其他方法改進。

1. 用nltk.word\_tokenize取代空白分字 <https://stackoverflow.com/questions/64675028/what-are-the-cases-where-nltks-word-tokenize-differs-from-str-split>
2. 用.strip()去除多餘空白
3. 用nltk.FreqDist 取代原本篩選的方法 <https://stackoverflow.com/questions/4634787/freqdist-with-nltk>
4. 由於改用Fasttext模型運算緩慢，因此改用伺服器，將原本的程式碼用GPT改成平行化的版本。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

* Training

Model 1 – Word2vec：一開始使用gensim的word2vec，但發現會有OOV的問題，在Slido看到同學有同樣的問題，而老師建議使用FastText，因此轉而使用FastText。

一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述參數設定：

vector\_size：詞向量的維度大小

window：上下文窗口，

min\_count：忽略低頻詞

max\_vocab\_size：詞彙表大小

Model2 – FastText：FastText會將詞語分解為子詞來訓練和生成詞向量，也就是可以處理從未見過的新詞，因此可以解決OOV問題。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述參數設定：

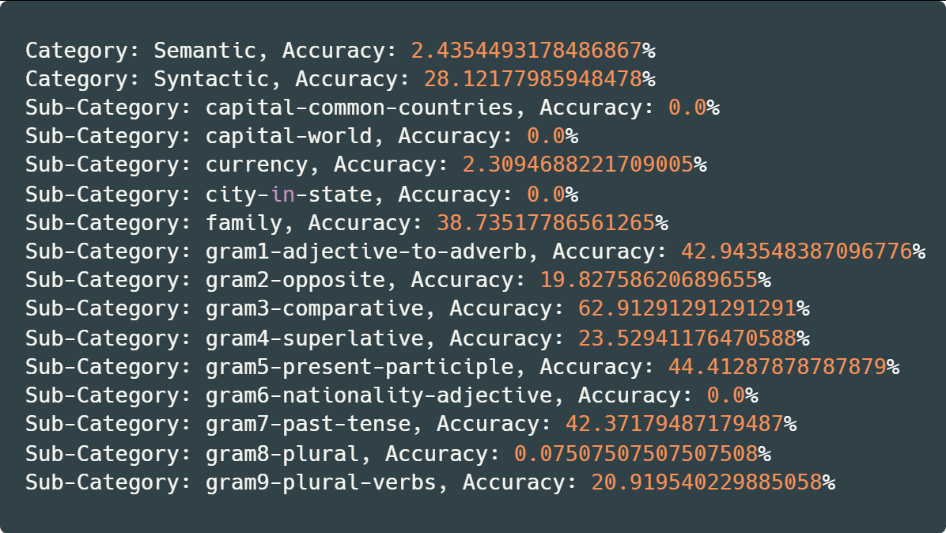
增加了vector\_size、window；

使用skip-gram模型；

sample：減少太常見的詞的影響；

negative：使用負採樣，加速模型訓練

Best Result：



對比於glove-wiki-gigaword-100和fasttext-wiki-news-subwords-300：

一張含有 文字, 螢幕擷取畫面, 字型, 功能表 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 功能表 的圖片

自動產生的描述

Semantic(國家首都、貨幣等)準確率非常低，特別是在capital-common-countries、capital-world 和 city-in-state 等子類別，準確率都為 0%。

句法類別的表現較好，尤其是 gram3-comparative（62.91%）和 gram5-present-participle（44.41%）子類別。

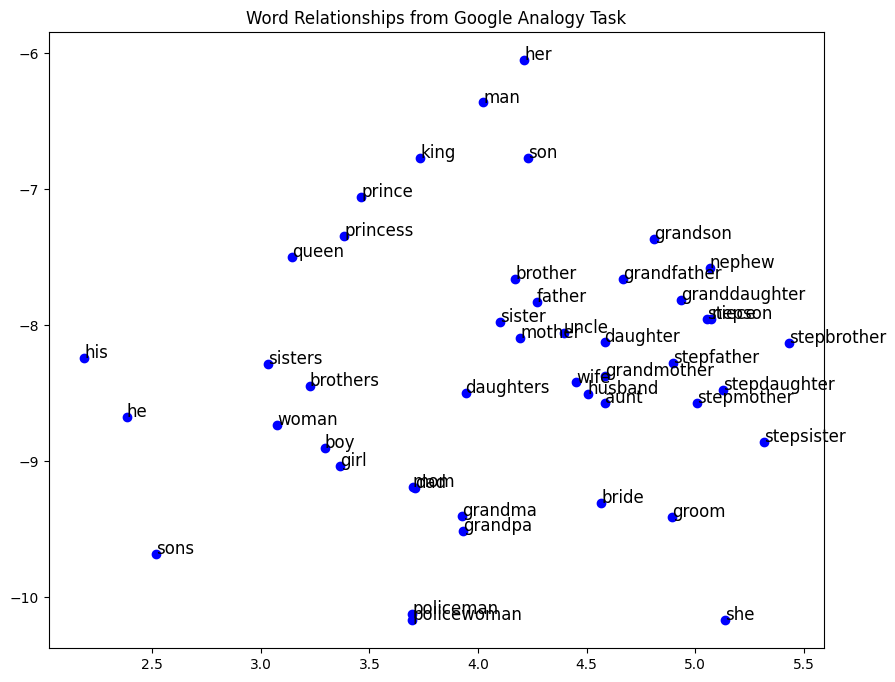
模型在語法相關的任務上表現較好，而在語義相關的任務上則有明顯不足，可以看出自己Train的Model離Pre-trained的Model的Accuracy還有一段差距。

另外，可以發現相對於glove-wiki-gigaword-100，fasttext-wiki-news-subwords-300有明顯較高的準確率，可能是由於：

1. vector\_size的大小，越多維度可以表示越多的東西。
2. fasttext如同前述，可以將詞語分解為子詞來訓練和生成詞向量，因此可以避免OOV的情況。

* t-SNE plot(TODO 7)：

由Plot出來的圖可以發現，相同類型(輩份關係)的詞大多會靠的較近，但由於模型準確度不夠，還是有許多相同輩分的男跟女相差很遠，例如daughters跟sons、woman跟man等。



相較之下，用fasttext-wiki-news-subwords-300 plot出來的圖就可以明顯感覺出不同關係的字的男女距離都非常靠近且相對距離很接近，因此在analogy task上可以表現很好。

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

* What’s the difference in word representations if you increase the amount of training data?

我嘗試使用了20%、25%、30%的訓練資料，測試結果發現準確率有小幅度提升，但沒有差非常多，可能是資料量差距不夠大，沒有使用更小的資料集的原因是模型準確度已經非常差了，減少資料量一定只會變更差。

另外，我也嘗試了許多模型參數的設置，但結果並沒有顯著差異。